Neuro-Oncology

Validation of neuroradiologic response assessment in gliomas: Measurement by RECIST, two-dimensional, computer-assisted tumor area, and computer-assisted tumor volume methods[1]

Evanthia Galanis,[2] Jan C. Buckner, Matthew J. Maurer, Rene Sykora, René Castillo, Karla V. Ballman, and Bradley J. Erickson for the North Central Cancer Treatment Group

Division of Medical Oncology (E.G., J.C.B.), Division of Biostatistics (M.J.M., K.V.B.), and Department of Radiology (R.S., B.J.E.), Mayo Clinic, Rochester, MN 55905; and Ochsner Clinic Foundation, New Orleans, LA 70121 (R.C.); USA

Significant limitations are associated with the use of standard radiographic measurements as indicators of response in glioma therapy trials. The Response Evaluation Criteria in Solid Tumors (RECIST) were recently introduced in an attempt to standardize and simplify assessment of response to treatment in cancer clinical trials. However, their applicability in gliomas has been assessed in only a very small number of patients. Our aim was to validate radiographic response assessment in newly diagnosed glioma patients. Sixty-seven newly diagnosed glioma patients participating in nine North Central Cancer Treatment Group glioma trials were included; 565 MRI scans were analyzed. All scans were performed with the same technique. Kappa statistics were calculated to determine agreement between assessment methods. Cox proportional hazards analyses and time-dependent Cox models were used to assess the association between different measurement methods and overall survival. Results showed agreement between the one-dimensional (1D) and two-dimensional (2D) measurements both for T2 images and for gadolinium-enhanced images. Comparison of duration of response and time to progression as assessed by eight different methods showed similarity in response assessments by 1D, 2D, area, and volume gadolinium measurements. In contrast, time to progression was significantly shorter when assessed by 1D–T2 or 2D–T2 images as compared to area–T2 or volume–T2 images. This set of data indicates that RECIST could be used instead of 2D imaging for response assessment in newly diagnosed glioma trials. Overall, responses as determined by any tumor measurement method did not correlate with patient survival for either enhancing or nonenhancing tumors, although the small number of responders limits definitive conclusions. Time-dependent Cox models demonstrated that, in contrast to the case of nonenhancing tumors, progression as determined by

Received April 5, 2005; accepted October 19, 2005.

[1] This work was supported by NIH grant CA84388 (E.G.) and NCCTG grant CA25224 (J.C.B.). Bradley J. Erickson is a member of the Siemens Medical Systems Advisory Board.

[2] Address correspondence to Evanthia Galanis, Division of Medical Oncology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905 (galanis.evanthia@mayo.edu).

[3] Abbreviations used are as follows: 1D, 2D, and 3D: one-, two-, and three-dimensional, respectively; CI, confidence interval; PROG, progression; RECIST, Response Evaluation Criteria in Solid Tumors; REGR, partial response; STAB, stable.

1D, 2D, area, and volume measurements in gadolinium-enhanced images was predictive of survival of patients with enhancing tumors. *Neuro-Oncology 8, 156–165, 2006 (Posted to Neuro-Oncology [serial online], Doc. 05-040, March 2, 2006. URL www.dukeupress.edu/neuro-oncology; DOI: 10.1215/15228517-2005-005)*

Keywords: gliomas, Response Evaluation Criteria in Solid Tumors (RECIST), tumor response, WHO criteria

There are significant limitations and methodological problems associated with use of radiographic response rates as end points in glioma trials. Glioma size measurements are frequently difficult to obtain because of lack of distinct borders, irregular shape, and variation in multiplanar dimensions. Furthermore, although measurements have been traditionally based on the enhancing part of the tumor, the area of T2 signal abnormality in MRI usually indicates infiltrating tumor cells and possibly edema and should be considered in response assessment. This is particularly important for nonenhancing tumors such as low-grade gliomas. Frequently, high response rates do not translate into benefits in progression-free or overall survival (Brada and Sharpe, 1996). Clearly, better methods to assess glioma response to treatment are needed.

Assessment of response to treatment in clinical trials has been traditionally performed by using the WHO criteria (Miller et al., 1981; WHO, 1979), with tumor size estimation based on bidimensional measurements. Several problems have been identified with use of WHO criteria, including variances among research groups in the minimum lesion size and number of lesions to be recorded, variability in the definition of progressive disease, and the need to incorporate newer technology such as CT or MRI three-dimensional (3D)[3] measurements into response assessment. The Response Evaluation Criteria in Solid Tumors (RECIST) criteria were the product of collaboration of WHO, the National Cancer Institute of the United States, the European Organization for Research and Treatment of Cancer, and the National Cancer Institute of Canada Clinical Trials Group; they are based on unidimensional tumor measurements and were recently introduced in an attempt to standardize and simplify assessment of response to treatment in cancer clinical trials (Therasse et al., 2000). The initial series of patients on whom the development of RECIST guidelines was based included breast, lung, ovary, melanoma, and sarcoma patients for the majority of cases, as well as 31 brain tumor patients from the National Cancer Institute of Canada Clinical Trial Group phase 2 and 3 trials (Therasse et al., 2000). Subsequent comparisons of RECIST and WHO response criteria in different patient populations appear to support their equivalence in common tumor types such as breast (Park et al., 2003; Prasad et al., 2003) and lung (Park et al., 2003). Nevertheless, for other tumor types with inadequate representation in the cohort of patients on whom development of RECIST guidelines was based, such as mesothelioma (Byrne and Nowak, 2004; Monetti et al., 2004) and pediatric tumors (McHugh and Kao, 2003), concerns have been expressed, and the issue has been raised of possible modification of the RECIST criteria in order to assess response more accurately. In a series of 32 pediatric patients (130 MRI scans), Warren et al. (2001) showed high concordance among 1D, 2D, and 3D methods in detecting partial response, but estimating time to disease progression appeared to be method dependent for childhood brain tumors. In this latter series, only 10 patients had high-grade gliomas. Given the small number of adult patients with brain tumors in the initial RECIST analysis, and the challenges associated with response assessment in primary brain tumors, there is a need for further comparative assessment of RECIST criteria prior to their routine incorporation into glioma trials and the neuro-oncology practice.

Our study compared the unidimensional RECIST criteria with the WHO bidimensional criteria as tools for assessing response in patients with newly diagnosed glioma. In addition, we investigated the value of incorporating computer-calculated area and volume measurements in the follow-up and assessment of response in this patient population.

Materials and Methods

MRI Technique

We identified 565 MRI studies performed on 67 patients that were enrolled into nine North Central Cancer Treatment Group newly diagnosed glioma protocols between 1991 and 2001. Patients were eligible to be included in this analysis if at the time of study enrollment they had measurable disease according to the RECIST definition (i.e., ⩾10-mm diameter) (Therasse et al., 2000). All imaging studies were performed with the same technique (5-mm fixed slices with 2.5-mm gap, with precontrast T1-weighted, postcontrast T1-weighted, conventional spin-echo proton-density, and T2-weighted images in the oblique-axial plane). All studies were de-identified to random numbers. The contours of the tumors were drawn on the T2/photon-density and postgadolinium T1 images by using semiautomated methods. The same neuroradiologist (B.J.E.) adjusted thresholds and seed points to define the outer margins of the tumor T2 image and gadolinium-enhanced images. The same threshold values were used for all slices of a particular sequence. These maps (both T1 and enhancement) were then analyzed to determine the major axis length on a slice (a 1D, or RECIST, measurement), the product of the major and minor axis (a 2D measurement), the greatest area of any single image (area), and the volume. *Major axis*, *minor axis*, *area*, and *volume* were defined as follows:

> *Major axis*: The longest diameter measured in the axial-oblique plane of acquisition, which is on a plane parallel to the anterior commissure–posterior commissure line. This represented the 1D (RECIST) measurement.

Minor axis: The greatest diameter perpendicular to the major axis in the axial-oblique plane of acquisition. This was multiplied by the major axis measure to provide the bidimensional (2D) measurement.

Area: The largest contiguous group of pixels on any slice. It was computed by finding the slice with the most pixels within the contour and multiplying this pixel value by the interpixel spacing in the X and Y directions.

Volume: The largest 3D contiguous group of pixels on any slice. It was computed by finding all the pixels of all slices within the contour and multiplying by the X, Y, and Z spacing.

Definition of Response

For all measurements, the patients were classified according to response status. Partial response (regression, REGR) was determined by comparison to the baseline scan, and progression (PROG) was determined by comparison to the prior scan with the smallest tumor measurement.

The following cutoffs were employed in order to define REGR or PROG and were based on the WHO (1979) and RECIST (Therasse et al., 2000) definitions of partial response and their correlation with volume (Therasse et al., 2000).

1D: REGR = –30%, PROG = +20%
2D: REGR = –50%, PROG = +25%
Area: REGR = –50%, PROG = +25%
Volume: REGR = –65%, PROG = +40%

Complete response was defined as complete disappearance of the patient's tumor. However, none of the 67 patients met radiographic criteria for complete response. Regression or complete response also required the patient to be on stable, decreased dose or off corticosteroids. Patients who did not meet the criteria for REGR or PROG were classified as stable (STAB).

Statistical Analysis

Categorical patient characteristics were summarized with the observed frequency and percent; age was summarized with the mean ± standard deviation as well as the median (minimum, maximum) age. Assessment of the amount of agreement among all distinct pairs of the eight measurements was summarized with observed frequency and percent, as well as with a weighted kappa statistic (Cohen, 1968) and its 95% confidence interval (CI). The kappa statistic measures the amount of agreement (i.e., correlation) between two measurements; a kappa value of 1 indicates perfect agreement, a value of 0 indicates lack of agreement, and a value of –1 indicates perfect disagreement. When the 95% CI for the kappa statistic does not contain 0, this indicates a statistically significant agreement between the two measures at the 0.05 level—that is, $P < 0.05$. The amount of agreement was determined for two variables: best objective response assessed at four months (PROG vs. STAB vs. REGR) and response/nonresponse at four months (REGR vs. STAB or PROG).

Outcome variables of interest were survival, progression-free survival, and duration of a response. In this study, all patients were part of clinical trials. No scans were available after a patient went off study because of progression. Survival was measured from time of study enrollment until death or last follow-up. Progression-free survival was measured from time of study enrollment until progression (as determined by the measurement method) or last follow-up. Duration of response was measured from time of an initial objective response of REGR until progression or last follow-up. All time-to-event measures were summarized with curves obtained from Kaplan–Meier estimates (Kaplan and Meier, 1958), and the median time and 95% CI for the median time are reported. Kaplan–Meier estimates for the time-to-event variables among the eight different tumor assessment measures were not directly compared because each measure was applied on the same set of patients.

The association between each tumor assessment method and patient outcome was evaluated by using Cox proportional hazards models. For this analysis, the response status at four months after starting study treatment was determined for each tumor assessment method. The four-month time point allows for evaluation of two eight-week treatment cycles; most observed antitumor activity will have occurred by this time. All patients in the study were alive at the four-month assessment; thus, a landmark analysis may be used for overall survival, where the response status is examined for predicting future survival (Hess et al., 1999). For each assessment method, comparisons to survival were made between patients classified as responders (REGR) and those classified as nonresponders (STAB or PROG) at the four-month time point, as well as between patients classified as progressors (PROG) and those who did not have progressive disease at four months (REGR or STAB). In addition, time-dependent Cox models were used to examine the relationship between survival and response (progression) status across the entire follow-up period. Specifically, patients changed from no response (or no progression) to response (or progression) at the time of the evaluation for which the tumor measurement satisfied the response (or progression) criterion; if a response (or progression) was not observed for a patient during follow-up, the patient remained in the no-response (no-progression) group throughout the follow-up period. All statistical tests were two-sided, and a P of <0.05 was considered to be statistically significant.

All analyses were done on the combined set of tumors, as well as on groups of tumors stratified by enhancement status (enhancing vs. nonenhancing) and by tumor grade (high vs. low). Grade 1 and 2 tumors were considered low grade, and grade 3 and 4 tumors were considered high grade.

Results

Patient Characteristics

All patients were treated as part of nine North Central Cancer Treatment Group glioma trials. They all received adjuvant radiation therapy after their initial surgery, and 85% also received nitrosourea-based chemotherapy. Table 1 summarizes patient characteristics by tumor enhancement (enhancing vs. nonenhancing) group. For the combined group, the median age of the patients was 40 years (range, 23–66 years); 70% were male. About half of the patients had a biopsy only (45%), and half had some degree of surgical resection (24% gross total resection and 28% subtotal resection). Twenty-seven patients (40%) had low-grade tumors (grade 1 or 2), and 40 patients (60%) had high-grade tumors; 36 patients (54%) had enhancing tumor, and 31 patients (46%) had nonenhancing tumors. The two groups differed with respect to age and tumor grade. Patients with enhancing tumors tended to be older ($P = 0.009$) and, as expected, had a higher grade tumor ($P = 0.04$) than patients with tumors that did not enhance.

Assessment of Agreement Among the Different Response Measures

To evaluate agreement among the different measurement methods in assessing the best objective response, we computed the number (and percent) of patients for each pair of methods in which there was agreement between the response assessments (REGR, STAB, or PROG) at four months, and we determined the distribution of response assessment at four months produced by each measurement method (Table 2). From Table 2, it can be seen that, for enhancing tumors, the measures on Gd-enhanced images were more likely to have a response classification of REGR than the corresponding measures on the T2 images. The area and volume measurements were more likely to produce a response of STAB than the 1D and 2D measures within a particular image type (T2 or Gd enhanced). For the nonenhancing tumors, area and volume measures were again more likely to produce a response status of STAB than the 1D and 2D measures; the 1D and 2D measures were more likely to have a response assessment of PROG.

Figure 1 shows the values of the weighted kappa estimates, and the corresponding 95% CIs, for the agreement between pairs of methods by response status (REGR, STAB, and PROG) at four months. In general, there was substantial agreement between the 1D and 2D measurements for both T2-enhanced and Gd-enhanced images. There was also strong agreement between area and volume measurements for both T2-enhanced and Gd-enhanced images. Although there was a statistically significant agreement between some pairs of measurements where one measure was on a T2-enhanced image and the other on a Gd-enhanced image, the agreement was generally weaker. Analyses on all tumors pooled and stratified by tumor grade yielded analogous results.

Patient Outcome Analysis

In general, relatively few patients had tumor responses by any of the tumor measurement techniques investigated (Table 3). Overall, there did not appear to be a significant association between response status and survival for any of the tumor measurement techniques for

Table 1. Patient characteristics

Variable	Enhancing Tumors (n = 36)	Nonenhancing Tumors (n = 31)	*P*
Age			
Mean ± SD	45 ± 11	38 ± 10	
Median (min, max)	45 (25, 66)	37 (23, 57)	0.009
Gender, n (%)			
Male	24 (67%)	23 (74%)	
Female	12 (33%)	8 (26%)	0.50
Extent of resection, n (%)			
Biopsy	14 (39%)	16 (52%)	
GTR	11 (31%)	5 (16%)	
STR	11 (31%)	8 (26%)	
Missing	0	2 (6%)	0.35
Tumor type/grade, n (%)			
Grade 1, 2 O or OA	7 (19%)	14 (45%)	
Grade 2 A	2 (6%)	4 (13%)	
AA or AOA	5 (14%)	4 (14%)	
GBM	22 (61%)	9 (29%)	0.04

Abbreviations: A, astrocytoma; AA, anaplastic astrocytoma; AOA, anaplastic oligoastrocytoma; GBM, glioblastoma multiforme; GTR, gross total resection; OA, oligoastrocytoma; STR, subtotal resection.

Table 2. Agreement among the measurement methods in terms of response (partial response, stable, and progression) at four months*

Measurement	Measurement Method							
Method	1D T2	2D T2	Area T2	Volume T2	1D Gd	2D Gd	Area Gd	Volume Gd
Enhancing tumors								
1D T2	36 (100%)	31 (86%)	19 (53%)	19 (53%)	19 (53%)	21 (58%)	18 (50%)	18 (50%)
2D T2		36 (100%)	22 (61%)	21 (58%)	22 (61%)	22 (61%)	20 (56%)	21 (58%)
Area T2			36 (100%)	27 (75%)	22 (61%)	23 (64%)	23 (64%)	23 (64%)
Volume T2				36 (100%)	22 (61%)	24 (67%)	26 (72%)	26 (72%)
1D Gd					36 (100%)	33 (92%)	26 (72%)	27 (75%
2D Gd						36 (100%)	25 (69%)	27 (75%)
Area Gd							36 (100%)	33 (92%)
Volume Gd								36 (100%)
REGR	4 (11%)	4 (11%)	2 (6%)	0 (0%)	6 (17%)	4 (11%)	6 (17%)	4 (11%)
STAB	8 (22%)	12 (33%)	17 (47%)	18 (50%)	14 (39%)	14 (39%)	16 (44%)	18 (50%)
PROG	24 (67%)	20 (56%)	17 (47%)	18 (50%)	16 (44%)	18 (50%)	14 (39%)	14 (39%)
Nonenhancing tumors								
1D T2	31 (100%)	28 (90%)	19 (61%)	16 (52%)				
2D T2		31 (100%)	16 (52%)	13 (42%)				
Area T2			31 (100%)	26 (84%)				
Volume T2				31 (100%)				
REGR	3 (10%)	4 (13%)	2 (6%)	1 (3%)				
STAB	15 (48%)	12 (39%)	25 (81%)	27 (87%)				
PROG	13 (42%)	15 (48%)	4 (13%)	3 (10%)				

Abbreviations: 1D, one-dimensional; 2D, two-dimensional; Gd, gadolinium enhanced; REGR, partial response (regression); STAB, stable; PROG, progression.

*Data represent the number of patients for whom comparison of measurements was made.

both the enhancing and nonenhancing tumors (Table 3). Only 2D Gd-enhancing tumors had a statistically significant association between response and survival when response status was treated as a time-dependent variable in the Cox model analysis ($P = 0.02$); the association was not significant when the response variable was the status at four months ($P = 0.76$). The lack of an observed significant association when using response status at four months as the variable in the Cox model could be explained by a lack of power—there are very few responses at four months across all tumor measurement techniques and across enhancing and nonenhancing tumors. While there were more tumor responses when using the time-dependent version of the response status variable, there still appeared to be a lack of association across all but one of the measurement methods and in both enhancing and nonenhancing tumors.

There were considerably more patients who experienced tumor progression as determined by the various tumor measurement techniques (Table 4) than there were patients who had tumor responses. For the enhancing tumors, the majority of the progressions occurred within four months. This was not the case for the nonenhancing tumors. Note that tumor progression as determined by 1D T2 and 2D T2 was found not to be significantly associated with survival for both enhancing and nonenhancing tumors. There was a statistically significant association between tumor progressions, as determined by 1D Gd, 2D Gd, area Gd, and volume Gd, and survival in the enhancing tumors; this was the case for both the progression status at four months and the time-dependent progression status variable. Individuals who were determined to have a progression on the basis of their tumor measurement had a worse survival. There also was an association between progression status at four months, as determined by area T2 and volume T2 measurements, and survival in nonenhancing tumors; however, this association was weaker and did not quite achieve statistical significance when progression status was treated as a time-dependent variable.

Determinations of Time to Response, Duration of Response, and Time to Progression

Estimates of the median time to response (i.e., to a tumor status assessment of REGR), the median duration of response (i.e., time from radiographic status of REGR to progression), and the median time to progression (i.e., time from study enrollment to progression) were determined for each of the eight measurements (Table 5). Area T2 and volume T2 tended to have a longer median time until a response was declared than all the other methods, regardless of image type and tumor enhancement status. The 1D and 2D median times to response also appeared similar within image type for

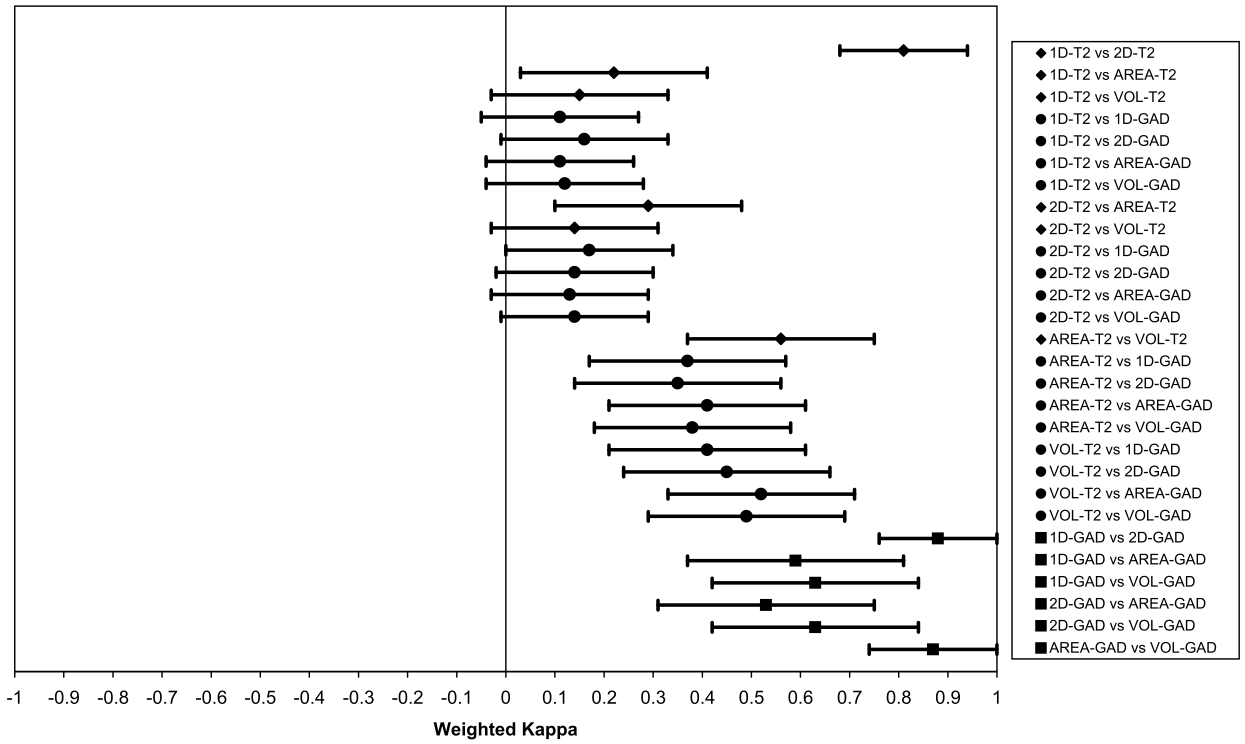

Fig. 1. Weighted kappa statistics value and corresponding 95% confidence interval (CI) for agreement in response assessment at four months (REGR, partial response; STAB, stable; and PROG, progression) between all distinct pairs of tumor assessment methods. A kappa value of 1 indicates a perfect agreement, whereas a value of –1 indicates perfect disagreement. When the 95% CI for kappa statistics does not contain 0, this indicates that there is statistically significant agreement between the two measurements at the 0.05 level.

both enhancing and nonenhancing tumors. In general, 1D and 2D measurements tended to have the shortest median response durations, regardless of image type and whether the tumor enhanced. Although the median duration of response appeared similar among all four measurements for nonenhancing tumors, this was not the case for the enhancing tumors. Finally, for the median time to progression, the volume and area measurements tended to have a longer median time to progression than the 1D and 2D measurements across image types and tumor enhancement status. This was most pronounced for the nonenhancing tumors, with the median time to progression for area T2 and volume T2 (15.8 and 47.3 months, respectively) being considerably longer than for the 1D T2 and 2D T2 measurements (2.5 and 4.1 months, respectively).

Discussion

Tumor size, both in clinical trials and in clinical practice, has traditionally been estimated from bidimensional measurements (the product of the longest diameter and its longest perpendicular diameter) (James et al., 1999), a methodology on which the WHO response criteria were based. However, measuring two tumor dimensions and then calculating their products and their sum can be laborious and associated with the risk of error. Changes in one diameter, however, should relate more closely to the fixed proportion of cells killed by the treatment than do changes in the bidimensional product (James et al., 1999; Therasse, 2002). Since the measurement methods and selection of target lesions were not clearly described in the WHO guidelines, assessment of tumor response has been shown to be poorly reproducible from one investigator or group of investigators to another (Therasse et al., 2000). Furthermore, the development of new imaging technologies and recent progress in the development of new classes of anticancer agents have required establishment of new methodology, which has led to a number of different modifications of WHO criteria (Padhani and Husband, 2000; Therasse, 2002). In 1998, a new set of response evaluation criteria for solid tumors—the RECIST criteria—were proposed by the RECIST working group in order to minimize the risk of measurement error and prevent the overestimation of response rates (Therasse et al., 2000). In neuro-oncology, the Macdonald criteria (Macdonald et al., 1990) have also been extensively used to assess response, especially in phase 2 glioma trials. They are based on WHO criteria; however, only changes in the enhancing part of the tumor are considered in assessing response and interpreted in conjunction with other parameters such as steroid use and neurologic findings.

When assessing response in gliomas, use of either WHO or RECIST response criteria is complicated by a

Table 3. Response predictor of overall survival

	Response at Four Months			Time-Dependent Cox Model		
Measurement Method	**No.**	**HR (95% CI)**	***P***	**No.**	**HR (95% CI)**	***P***
Enhancing tumors						
1D T2	4	2.27 (0.76, 6.75)	0.14	10	1.30 (0.62, 2.73)	0.49
2D T2	4	1.74 (0.59, 5.11)	0.31	12	1.04 (0.50, 2.13)	0.93
Area T2	2	1.69 (0.40, 7.23	0.48	9	0.61 (0.25, 1.49)	0.28
Volume T2	0	—	—	6	0.48 (0.16, 1.43)	0.19
1D Gd	6	0.60 (0.18, 2.02)	0.41	16	0.77 (0.39, 1.55)	0.47
2D Gd	4	0.80 (0.19, 3.41)	0.76	16	0.36 (0.15, 0.86)	0.02
Area Gd	6	0.91 (0.31, 2.66)	0.86	14	0.63 (0.32, 1.27)	0.20
Volume Gd	4	1.44 (0.43, 4.87)	0.56	11	1.07 (0.52, 2.22)	0.85
Nonenhancing tumors						
1D T2	3	1.49 (0.34, 6.61)	0.60	9	0.67 (0.28, 1.59)	0.36
2D T2	4	1.91 (0.53, 6.86)	0.32	8	0.62 (0.26, 1.49)	0.29
Area T2	2	1.29 (0.17, 9.96)	0.81	8	1.26 (0.55, 2.87)	0.59
Volume T2	1	2.19 (0.28, 16.96	0.45	6	1.08 (0.43, 2.71)	0.88

Abbreviations: 1D, one-dimensional; 2D, two-dimensional; CI, confidence interval; Gd, gadolinium enhanced; HR, hazard ratio.

Table 4. Progression as predictor of overall survival

	Progression Status at Four Months			Time-Dependent Cox Model		
Measurement Method	**No.**	**HR (95% CI)**	***P***	**No.**	**HR (95% CI)**	***P***
Enhancing tumors						
1D T2	4	2.27 (0.76, 6.75)	0.14	10	1.30 (0.62, 2.73)	0.49
1D T2	24	1.48 (0.63–0.3.45)	0.37	34	1.54 (0.36–6.68)	0.56
2D T2	20	1.32 (0.60–2.93)	0.49	33	1.79 (0.53–6.01)	0.35
Area T2	17	3.64 (1.55–8.55)	0.003	28	1.93 (0.86–4.32)	0.11
Volume T2	18	6.09 (2.44–15.22)	0.0001	28	1.87 (0.84–4.19)	0.13
1D Gd	16	4.42 (1.83–10.66)	0.0009	22	2.39 (1.18–4.81)	0.02
2D Gd	18	4.68 (1.87–11.71)	0.001	23	4.27 (1.85–9.85)	0.0007
Area Gd	14	11.06 (3.74–32.66)	<0.0001	24	2.36 (1.14–4.86)	0.02
Volume Gd	14	4.71 (1.99–11.13)	0.0004	24	2.03 (1.00–4.15)	0.05
Nonenhancing tumors						
1D T2	13	1.23 (0.45–3.33)	0.69	29	0.30 (0.08–1.14)	0.08
2D T2	16	1.86 (0.68–5.07)	0.23	27	0.96 (0.20–4.64)	0.96
Area T2	4	5.14 (1.54–17.20)	0.008	26	1.96 (0.68–5.63)	0.21
Volume T2	3	3.55 (0.77–16.40)	0.10	17	1.33 (0.62–2.85)	0.47

Abbreviations: 1D, one-dimensional; 2D, two-dimensional; CI, confidence interval; Gd, gadolinium enhanced; HR, hazard ratio.

more fundamental question, which is whether conventional oncological criteria of response when translated into CNS tumors represent a useful measure and a true reflection of treatment efficacy. The poor correlation between response as measured in phase 2 studies and survival in adjuvant studies suggests that there may be methodological flaws. Within the brain, a reduction in the size of an enhancing abnormality may represent either loss of tumor cells or other processes such as an alteration in the properties of the blood–brain barrier. Even if decreased size is indicative of tumor cell death, the assessment of radiological response is difficult. Although agreement on response definition provides a common language, it is not always clear whether it is accurately associated with the principal end point: survival. The goal of this study was to address some of these issues by comparing 1D, 2D, area, and volume measurements in patients with newly diagnosed glioma and correlate them with outcome.

In our study, RECIST 1D measurements were com-

Table 5. Summaries of time to response, duration of response, and time to progression

	Median Number of Months (95% CI)		
Measurement Method	**Time to Response**	**Duration of Response**	**Time to Progression**
Enhancing tumors			
1D T2	1.6 (1.1–2.3)	5.6 (1.6–10.1)	2.5 (1.9–4.0)
2D T2	2.5 (1.2–4.0)	3.8 (2.2–9.2)	3.0 (1.9–5.7)
Area T2	15.1 (3.7–20.3)	NA (9.5–NA)	4.9 (3.7–8.7)
Volume T2	12.3 (2.9–20.8)	7.2 (2.2–NA)	4.8 (3.2–8.9)
1D Gd	3.8 (1.4–6.2)	NA (1.8–NA)	4.9 (2.3–NA)
2D Gd	5.3 (3.7–6.2)	NA (2.2–NA)	3.6 (1.9–NA)
Area Gd	4.0 (1.8–7.9)	8.7 (3.5–NA)	5.8 (3.8–18.4)
Volume Gd	4.1 (1.7–9.7)	13.6 (2.8–NA)	5.8 (3.8–27.3)
Nonenhancing tumors			
1D T2	5.0 (1.5–7.1)	9.7 (3.2–NA)	5.8 (3.7–8.8)
2D T2	4.5 (1.4–18.4)	12.5 (3.2–24.7)	4.1 (3.0–7.4)
Area T2	11.5 (1.9–20.7)	12.7 (2.4–NA)	15.8 (8.8–29.7)
Volume T2	29.3 (11.0–38.8)	12.4 (4.7–NA)	47.3 (15.7–NA)

Abbreviations: 1D, one-dimensional; 2D, two-dimensional; CI, confidence interval; HR, hazard ratio; NA, not achieved

parable to 2D measurements in determining time to response, duration of response, and time to progression (Table 5). Furthermore, there was agreement between RECIST and 2D measurements both in Gd-enhanced and T2 images (Fig. 1; kappa = 0.87 [CI, 0.73–1.00] and kappa = 0.81 [CI, 0.68–0.94], respectively). These data are consistent with the comparative analysis of the two methodologies performed on 30 brain tumor patients, which was also taken into account for the development of RECIST guidelines, and they support replacement of 2D with RECIST response criteria in neuro-oncology clinical trials.

Although there was good agreement between RECIST 1D and 2D measurements, as Table 2 indicates (86% and 90% for enhancing and nonenhancing tumors, respectively), as shown in Tables 3 and 4, neither measurement appears to predict outcome. Specifically, no association was found between response, as assessed by these two methods, and survival. In this respect, neither method appears superior to the other. Nevertheless, the small number of responders could have significantly decreased the likelihood of identifying existing associations.

As it pertains to volumetric measurements, there was good agreement between volume and 1D and 2D measurements in Gd-enhanced images (75% agreement with both 1D and 2D measurements [Table 2]). The agreement was much weaker, however, in the T2 images (Table 2 and Fig. 1). Furthermore, according to this set of data, neither Gd nor T2 volume measurements appeared to predict outcome for either enhancing or nonenhancing tumors.

In our data set, response at four months was not predictive of overall survival for any assessment method (all $P > 0.14$), while the only significant association between response and survival in time-dependent Cox models pertained to 2D measurements in Gd-enhanced images ($P = 0.02$). The small number of responders prevents definitive conclusions, however. In contrast, the 1D Gd–measured and the 2D Gd–measured progression at four months was predictive of overall survival ($P = 0.0009$ and 0.001, respectively). There was no such association for 1D T2 and 2D T2 measurements, however ($P > 0.23$ for all enhancing and nonenhancing tumors [Tables 3 and 4]).

These results emphasize significant methodological problems associated with assessment of response of nonenhancing tumors such as low-grade gliomas to treatment: Responses based on 2D T2 images do not associate well with patient outcome, and 1D (RECIST) T2 images fare equally poorly. There is an important need to incorporate and prospectively validate imaging methodology that can better predict outcome of nonenhancing tumors in low-grade glioma trials.

The other point that these data emphasize is that when time to progression is used as the primary outcome, results may vary widely, depending on the imaging methodology used. This is illustrated by the shorter time to progression of nonenhancing tumors when assessment is performed by 1D T2 or 2D T2 images as compared to area T2 or volume T2 images. It is particularly pertinent for low-grade tumors, in which T2 measurements represent the mainstay for assessment of treatment efficacy. In contrast, 1D, 2D, area, and volume Gd measurements perform similarly with regard to duration of response and time to progression.

A frequent concern when bidimensional or unidimensional measurements are employed on imaging studies pertains to intraobserver and interobserver variability. This can be quite high (Hopper et al., 1996; Lavin and Flowerdew, 1980; Quoix et al., 1988; Thiesse et al., 1997; Warr et al., 1984), presumably because of the subjectivity involved in defining the exact margins of the lesion and determining the lesion's largest diameter and its largest perpendicular diameter (Fornage, 1993). Such variability can have a significant impact on the assessment of an individual patient's tumor response to a given therapy, as well as the determination of the efficacy of a new antitumor therapy (Lavin and Flowerdew, 1980; Thiesse et al., 1997; Warr et al., 1984). Schwartz and coworkers (2000) have shown that tumor size can be obtained more accurately and consistently by readers using an automated autocontour technique than by those using handheld or electronic calipers. Autocontouring in their series was performed with the radiologist placing a cursor in the center of the lesion and the computer determining the border of the lesion on the basis of density differences. In our set of data, the area-T2 and area-Gd determinations were based on computer determination of the tumor area, which was based on the largest contiguous group of pixels on any slide. It is of note that in our study, both for enhancing and for nonenhancing tumors, a progression status that was defined by computer-calculated tumor area or volume measurements in T2 images at four months performed significantly better in predicting survival than did a progression status that was defined by 1D T2 or 2D T2 measurements (Tables 3 and 4). A possible explanation for this could be the higher sensitivity of the density-based area determination approach in assessing the real extent of the lesion in the absence of enhancement.

Our series includes only patients with newly diagnosed glioma. Although our conclusions could also be applicable for patients with recurrent glioma, additional methodological difficulties apply, especially when assessing the response to newer treatment modalities, such as biologics or molecular targeted therapies. The value of RECIST versus 2D measurements versus the added value of other methodology, that is, area-based or volume-based determinations of response, versus use of functional imaging such as thallium 201 single-proton-emission computer tomography (Vos et al., 2003) in assessing response to treatment in patients with recurrent/progressive disease will need to be further evaluated. We are currently performing such an analysis.

In summary, our analysis results support the conclusion that RECIST could be used instead of conventional 2D imaging in trials with patients who have newly diagnosed glioma. Overall responses as determined by any tumor measurement method did not correlate with patient survival for either enhancing or nonenhancing tumors, although the small number of responders limits definitive conclusions. In time-dependent Cox models, progression as determined by 1D, 2D, area, and volume measurements in Gd-enhanced images was predictive of survival of patients with enhancing tumors.

References

Brada, M., and Sharpe, G. (1996) Chemotherapy of high-grade gliomas: Beginning a new era or the end of the old? *Eur. J. Cancer* **32A**, 2193–2194.

Byrne, M.J., and Nowak, A.K. (2004) Modified RECIST criteria for assessment of response in malignant pleural mesothelioma. *Ann. Oncol.* **15**, 257–260.

Cohen, J. (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **70**, 213–220.

Fornage, B.D. (1993) Measuring masses on cross-sectional images. *Radiology* **187**, 289 (letter).

Hess, K.R., Wong, E.T., Jaeckle, K.A., Kyritsis, A.P., Levin, V.A., Prados, M.D., and Yung, W.K.A. (1999) Response and progression in recurrent malignant glioma. *Neuro-Oncology* **1**, 282–288.

Hopper, K.D., Kasales, C.J., Van Slyke, M.A., Schwartz, T.A., TenHave, T.R., and Jozefiak, J.A. (1996) Analysis of interobserver and intraobserver variability in CT tumor measurements. *AJR. Am. J. Roentgenol.* **167**, 851–854.

James, K., Eisenhauer, E., Christian, M., Terenziani, M., Vena, D., Muldal, A., and Therasse, P. (1999) Measuring response in solid tumors: Unidimensional versus bidimensional measurement. *J. Natl. Cancer Inst.* **91**, 523–528.

Kaplan, E.L., and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481.

Lavin, P.T., and Flowerdew, G. (1980) Studies in variation associated with the measurement of solid tumors. *Cancer* **46**, 1286–1290.

Macdonald, D.R., Cascino, T.L., Schold, S.C., Jr., and Cairncross, J.G. (1990) Response criteria for phase II studies of supratentorial malignant glioma. *J. Clin. Oncol.* **8**, 1277–1280.

McHugh, K., and Kao, S. (2003) Response evaluation criteria in solid tumours (RECIST): Problems and need for modifications in paediatric oncology? *Br. J. Radiol.* **76**, 433–436.

Miller, A.B., Hoogstraten, B., Staquet, M., and Winkler, A. (1981) Reporting results of cancer treatment. *Cancer* **47**, 207–214.

Monetti, F., Casanova, S., Grass, A., Cafferata, M.A., Ardizzoni, A., and Neumaier, C.E. (2004) Inadequacy of the new Response Evaluation Criteria in Solid Tumors (RECIST) in patients with malignant pleural mesothelioma: Report of four cases. *Lung Cancer* **43**, 71–74.

Padhani, A.R., and Husband, J.E. (2000) Commentary: Are current tumour response criteria relevant for the 21st century? *Br. J. Radiol.* **73**, 1031–1033.

Park, J.O., Lee, S.I., Song, S.Y., Kim, K., Kim, W.S., Jung, C.W., Park, Y.S., Im, Y.H., Kang, W.K., Lee, M.H., Lee, K.S., and Park, K. (2003) Measuring response in solid tumors: Comparison of RECIST and WHO response criteria. *Jpn. J. Clin. Oncol.* **33**, 533–537.

Prasad, S.R., Saini, S., Sumner, J.E, Hahn, P.F., Sahani, D., and Boland, G.W. (2003) Radiological measurement of breast cancer metastases to lung and liver: Comparison between WHO (bidimensional) and RECIST (unidimensional) guidelines. *J. Comput. Assist. Tomogr.* **27**, 380–384.

Quoix, E., Wolkove, N., Hanley, J., and Kreisman, H. (1988) Problems in radiographic estimation of response to chemotherapy and radiotherapy in small cell lung cancer. *Cancer* **62**, 489–493.

Schwartz, L.H., Ginsberg, M.S., DeCorato, D., Rothenberg, L.N., Einstein, S., Kijewski, P., and Panicek, D.M. (2000) Evaluation of tumor measurements in oncology: Use of film-based and electronic techniques. *J. Clin. Oncol.* **18**, 2179–2184.

Therasse, P. (2002) Measuring the clinical response. What does it mean? [erratum in *Eur. J. Cancer* [2003] **39**, 1489] *Eur. J. Cancer* **38**, 1817–1823.

Therasse, P., Arbuck, S.G., Eisenhauer, E.A., Wanders, J., Kaplan, R.S., Rubinstein, L., Verweij, J., Van Glabbeke, M., van Oosterom, A.T., Christian, M.C., and Gwyther, S.G. (2000) New guidelines to evaluate the response to treatment in solid tumors. *J. Natl. Cancer Inst.* **92**, 205–216.

Thiesse, P., Ollivier, L., Di Stefano-Louineau, D., Negrier, S., Savary, J., Pignard, K., Lasset, C., and Escudier, B. (1997) Response rate accuracy in oncology trials: Reasons for interobserver variability. *J. Clin. Oncol.* **15**, 3507–3514.

Vos, M.J., Hoekstra, O.S., Barkhof, F., Berkhof, J. Heimans, J.J., van Groeningen, C.J., Vandertop, W.P., Slotman, B.J., and Postma, T.J. (2003) Thallium-201 single-photon emission computed tomography as an early predictor of outcome in recurrent glioma. *J. Clin. Oncol.* **21**, 3559–3565.

Warr, D., McKinney, S., and Tannock, I. (1984) Influence of measurement error on assessment of response to anticancer chemotherapy: Proposal for new criteria on tumor response. *J. Clin. Oncol.* **2**, 1040–1046.

Warren, K.E., Patronas, K., Aikin, A.A., Albert, P.S., and Balis, F.M. (2001) Comparison of one-, two-, and three-dimensional measurements of childhood brain tumors. *J. Natl. Cancer Inst.* **93**, 1401–1405

WHO, World Health Organization (1979) *Handbook for Reporting Results of Cancer Treatment* (WHO Offset Publication No. 48). Geneva: World Health Organization.